(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
28 November 2002 (28.11.2002)

PCT

(10) International Publication Number
WO 02/095730 A1

(51) International Patent Classification[7]: G10L 15/20

(21) International Application Number: PCT/GB02/02197

(22) International Filing Date: 20 May 2002 (20.05.2002)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
0112319.9 21 May 2001 (21.05.2001) GB

(71) Applicant (*for all designated States except US*): QUEEN'S UNIVERSITY OF BELFAST [GB/GB]; 8 Malone Road, Belfast BT9 5BN (GB).

(72) Inventor; and
(75) Inventor/Applicant (*for US only*): MING, Ji [GB/GB]; 10 Rossdale Heights, Belfast BT8 6XZ (GB).

(74) Agent: MURGITROYD & COMPANY; Scotland House, 165-169 Scotland Street, Glasgow G5 8PL (GB).

(81) Designated States (*national*): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZM, ZW.

(84) Designated States (*regional*): ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

WO 02/095730 A1

(54) Title: INTERPRETATION OF FEATURES FOR SIGNAL PROCESSING AND PATTERN RECOGNITION

(57) Abstract: A method of interpretation of features for signal processing and pattern recognition provides a model in which the pattern or signal to be interpreted is considered as a set of N observations, M of which are corrupt, and a disjunction is performed over all possible combinations of N different values (1,...,N) taken N-M at a time. The value of M defines the order of the model, and is determined using an optimality criterion which chooses the order that corresponds to a clean signal based on comparing the state duration probability of the signal or pattern to be interpreted with that of a clean signal.

Interpretation of Features for Signal Processing and Pattern Recognition

The present invention relates to interpretation of features for signal processing and pattern recognition, and particularly to speech recognition subjected to partial, unknown frequency-based corruption.

Partial frequency-band corruption may account for the effect of a family of real-world noises, for example, a telephone ring, a car horn, a siren or a random channel tone, which usually have a band-selective characteristic and thus affect only certain parts of the speech frequency band. There may be two different ways to deal with this type of noise corruption for robust speech recognition. Firstly, we may use the conventional noise filtering or feature/model compensation techniques to remove the noise component from the input signal, or to adapt the model to the noisy environment. Each of these techniques assumes the availability of certain

knowledge of the noise or environment. The required knowledge may include, for example, the spectral or cepstral characteristics of the noise for noise filtering or feature selection, a stochastic model of the noise for noise compensation and an extra set of training data in the new environment for model adaptation.

The second possible way of dealing with this partial corruption is to base the recognition mainly on information from the clean frequency bands, by throwing away the noisy bands, or by making these bands play a less significant role in recognition, i.e., the missing feature method.

This recognition is made possible due to redundancy in the spectral characteristics of speech. This method is of interest because there can be situations where removing the noise from the input signal may prove difficult, due to the lack of sufficient knowledge about the noise. This lack of knowledge may be experienced, for example, when an unknown unexpected noise occurs in the middle of utterance. A better system may be a combination of these two methods, i.e., using the noise reduction technique to remove the noise with a known or stationary characteristic, and exploiting the redundancy in the speech signal to get around the noise with an unknown or time-varying nature. The present invention focuses on the second method, but we use a simple example to demonstrate the advantage of combining the two methods. In particular, we

study the sub-band approach for speech recognition involving partial unknown frequency-band corruption.

As a system paradigm for dealing with partial frequency-band corruption, the sub-band based approach has aroused much research interest over the past years. In this approach, the full speech frequency band is divided into several sub-bands, and each sub-band is featured independently of the other sub-bands, so that the local distortions in the frequency band will not spread over the entire feature space. Therefore, instead of requiring a detailed knowledge of the noise for clearing the corrupted sub-band features, the sub-band method, and in general the missing feature methods, require only a labelling of every sub-band/feature as reliable or corrupt, for removing the unreliable features from recognition.

Unfortunately, locating the corrupted sub-bands itself can be a difficult task, if there is no prior information on the noise. Mistakes in labelling the sub-bands can cause either a loss of reliable information, or an inclusion of unreliable information in the recognition process. This problem, i.e. extracting reliable features from a sub-band observations while assuming no prior knowledge on the noise, has been referred to as sub-band combination.

Recent studies have suggested several methods. Typically, these include the weighted-average

method, the neural-network method and the full-combination method.

In the weighted-average method, the likelihood from the individual sub-bands are combined by using a geometric or arithmetic average; the contribution of each sub-band is weighted by the local signal-to-noise ratio (SNR) related to that sub-band.

In the neutral-net method, independent networks are trained to estimate the probabilities of all possible combinations of subsets of the sub-bands, assuming that there exists at least one combination that accounts for the clean speech. This method faces the problem of how to select the best combination from all the combinations given no knowledge about the noisy bands. Some heuristic methods, such as majority voting or distance pruning, have been studied for this purpose.

The idea of explicitly creating all possible combinations among the sub-bands has been further studied in the full-combination model, in which the likelihood of different combinations of different sub-bands are combined using a weighted-average method, with each weight proportional to the relative reliability of a specific set of sub-bands.

In addition, the mixture of experts theory has also been discussed as a possible means of sub-band combination.

Clearly, a reliable estimation of the local noise characteristic or SNR is crucial to the success of the weighted-average model and full-combination model. In fact, it is crucial to the success of all missing feature methods which rely on an accurate mask for labelling the reliable and corrupt regions over the temporal-spectral feature space. The local SNR at each time-frequency location may be estimated by using the traditional spectral estimation approach, involving a running estimate of the local noise spectrum via spectral subtraction. This method performs well when the corrupting noise is stationary. But it may fail to produce accurate estimates in non-stationary noise or unknown noise, as in these conditions the assumption required for spectral subtraction is invalidated. To overcome this problem, it has been suggested that some characteristics of the speech signal itself, such as the harmonic nature of voiced speech may be exploited for identifying the corrupted time-frequency regions.

According to the present invention there is provided a method of interpreting features for signal processing and pattern recognition as described in the attached Claims.

The present invention proposed a new approach, the probabilistic union model, for combining the sub-band features with unknown, time-varying partial corruption. Unlike the missing feature method, the new model does not require the identity of the

corrupted bands, instead, it combines the sub-band features based on the probability theory for the union of random events, to account for any possible partial corruption with the sub-bands. This model improves upon the previous methods in that it offers robustness against partial frequency-band corruption, while requiring little or no information about the noise. We have incorporated the new union model into an HMM framework and tested it on a number of isolated word databases. The results have indicated the advantage of the union model over the previous methods for sub-band combination, particularly for dealing with band-selective noise with an unknown or time varying band location and/or bandwidth.

The present invention will now be described by way of example only, with reference to the accompanying tables and drawings in which;

Tables I and II show experimental results showing the performance of first and second embodiments of the present invention against a conventional technique, for incorrupt and corrupt signals respectively;

Tables III and IV show experimental results showing the performance of the second embodiment of the present invention against further conventional techniques, for stationary and non-stationary corruption respectively;

Table V shows experimental results showing the performance of a third embodiment of the present invention in comparison to the first and second embodiments and further conventional techniques;

Table VI shows experimental results showing the performance of the third embodiment of the present invention in comparison to the second embodiment;

Table VII shows experimental results showing the performance of a fourth embodiment of the present invention;

Fig. 1 illustrates the performance of a specific aspect of the present invention, and;
Fig. 2 illustrates the raw data used to test the performance of the present invention.

PROBABILISTIC UNION MODEL

A. Background

Assume a recognition system with N sub-bands, in which a speech utterance may be represented by N sub-band feature streams o_1, o_2, ..., o_N, where o_n represent the feature stream from the n'th sub-band. The presence of a band-selective noise can cause some of the o_n's to be corrupted. Thus, in recognition we face the problem of how to extract information for the utterance from a sub-band feature set $\{o_1, o_2, \cdots, o_N\}$, in which some of the

sub-band features o_n's may be noisy, but without knowledge about their identity.

When there is no noise the traditional approach for extracting the information is to combine the sub-band features by using the "and" (i.e. conjunction) operator ∧ (although this is not usually explicitly stated), i.e.

$$O_\wedge = o_1 \wedge o_2 \wedge \ldots \wedge o_N \tag{1}$$

where $O\wedge$ represents the combined observation. Assuming that the sub-band features are independent of one another, then the likelihood of $O\wedge$, $p(O\wedge)$, equals the product of the individual sub-band likelihoods $p(o_n)$'s i.e.

$$\begin{aligned} p(O\wedge) &= p(o_1 \wedge o_2 \wedge \ldots \wedge o_N) \\ &= p(o_1)p(o_2)\ldots p(o_N) \end{aligned} \tag{2}$$

For convenience, we call (1) the *product model*. Assume that the model, consisting of the probability densities of the individual sub-bands, $P(x_n)$'s is trained on clean speech to maximise the likelihood of some clean utterances. When this model is used for an utterance with some noisy sub-bands, then the corresponding $P(o_n)$'s for the noisy o_n's will become problematic, especially if the noise is strong. Typically, these noisy likelihoods may become very small on the correct model because of the poor match between the model and data. These small and random

sub-band likelihoods may easily dominate the product, and then destroy the model's ability to produce high likelihoods for correct phonetic classes. Simply removing the sub-band likelihoods with small values from the models may not improve this, because low likelihoods may also be the result of a phonetic mismatch, and because the likelihoods corresponding to the noisy sub-bands may not be small on the incorrect models which accidentally match the noisy data. This problem can be improved if the noisy sub-bands can be identified, whereby the corresponding likelihoods can be removed or "integrated" from the product, i.e. the missing feature method. This identification requires the local SNR related to each sub-band. This information may not be available for applications involving unknown, time-varying noise. This problem has been addressed by using a back-off model, in which each observation probability density is formed as a weighted combination of two densities: one from the training data and another, a uniform distribution, to account for possible outliers arising from the noise.

In the following we describe the probabilistic union model as an alternative, to overcome the above mentioned problems. We start to describe the model without considering the number of noisy sub-bands (except that the corruption is partial within the sub-bands); then we move to an extended model which takes into account knowledge on the number of noisy sub-bands.

B. General union model

Given no knowledge about the noisy sub-bands, we can alternatively assume that, in a given set of sub-band features $\{o_1, o_2, \cdots, o_N\}$, the reliable features that characterize the speech utterance may be *any* of the o_n's $n = 1, \ldots, N$, or *any* of the combinations among the o_n's up to the complete feature set. This can be expressed, using the inclusive "or" (i.e. disjunction) operator $\vee$, as

$$O\vee = o_1 \vee o_2 \vee \ldots \vee o_N$$

$$= \bigvee_{n=1}^{N} o_n \qquad (3)$$

where $O_\vee$ is a combined observation based on $\vee$, representing the reliable features within $\{o_1, o_2, \ldots o_N\}$.

For example, using a 3-band model, the expression $O\vee = o_1 \vee o_2 \vee o_3$ based on inclusive "or" assumes that the reliable features within the given $\{o_1, o_2, o_3\}$ may be $o_1, or\ o_2, or\ o_3, or\ o_1 \wedge o_2$, or $o_1, \wedge o_3, or\ o_2 \wedge o_3, or\ o_1 \wedge o_2 \wedge o_3$. These feature combinations can characterize, respectively, a speech utterance in which there are two-band, one-band and no band corruption, therefore covering all possible partial corruptions, including

the no corruption case which may be encountered in a 3-band system. In general, if an observation consists of N features $o_1, o_2, \cdots, o_N$, and these features may be subjected to some partial corruption with unknown characteristics, i.e. number and location of the corrupted features and statistics of the corrupting noise, then the useful information contained in the observation may be modelled by (3). This model takes into account all possible partial corruptions, thereby requiring no knowledge on the actual corrupting noise.

If we assume that the o_n's are discrete random vectors, then $O\vee$ is the union of the random events o_n's. Thus, we can compute the probability $P(O\vee)$ based on the rules of probability for the union of random events. This probability, for each modeled phonetic class, can then be used to decide the recognition result based on the maximum-likelihood principle. Note that $\vee_{n=1}^{m} o_n = \left(\vee_{n=1}^{m-1} o_n\right) \vee o_m$, so $P(O\vee)$ can be computed using a recursion

$$P\left(\vee_{n=1}^{m} o_n\right) = P\left(\vee_{n=1}^{m-1} o_n\right) + P(o_m) - P\left(\left(\vee_{n=1}^{m-1} o_n\right) \wedge o_m\right) \tag{4}$$

for m=2, . . . , N. With the assumption that the o_n's are mutually independent, then (4) can be simplified as

$$P(\vee_{n=1}^{m} o_n) = P(\vee_{n=1}^{m-1} o_n) + P(o_m) - P(\vee_{n=1}^{m-1} o_n) P(o_m) \tag{5}$$

This computation requires only the probability distributions of the individual sub-bands, i.e. $P(x_n)$'s which are assumed to be estimated from *clean* training data. We call (3)-(5) the *probabilistic-union model*, which extracts information based on the union of events. This is opposed to the product model (1)-(2), which extracts information based on the intersection of events.

Since the $P(o_n)$'s are generally not large, (5) is effectively the sum of the individual sub-band probabilities. A major difference between (5) and (2) (i.e. the product model) is that a small $P(o_n)$ makes only a small contribution to (5). Therefore a noisy sub-band, typically with low probability on the correct model, will have little effect on the union probability $P(O\vee)$ associated with the correct model. In other words, the union probability $P(O\vee)$ associated with the correct model is dominated by noiseless sub-bands, unlike the product model in which the likelihood associated with the correct model may be dominated by those small, random and noisy sub-band likelihoods. This effectively increases the probability associated with the correct model, such that, as long as the remaining clean sub-bands contain sufficient discriminative information, the correct model should still be able to score highly among the competitive models.

However, (5) has a disadvantage, i.e., it effectively averages the ability of each sub-band to

discriminate between correct and incorrect phonetic classes, unlike the product model in which each sub-band reinforces the other as the joint probability of the sub-band features is modeled.

This characteristic makes (5) an ineffective model both for utterances with more than one clean sub-band, and for clean utterance without band corruption. This problem may be overcome by combining the use of "and" and "or" operators, assuming a knowledge on the number of corrupted sub-bands. This is the extended union model described below.

C. Extended Union Model

In a first embodiment of the present invention, we aim to include all the clean sub-band features into a conjunction (i.e. combining them using the "and" operator), such that a joint probability of the clean features can be derived, which should be more powerful than any of their marginal probabilities in terms of discrimination. This can be achieved by combining the use of "and" and "or" operators, assuming only a knowledge on the number (not the location) of corrupted sub-bands. Specifically, for a given set of sub-band features $\{o_1, o_2, ..., o_N\}$ if the number of corrupted bands is M (M < N), then we know that there exists one subset of (N - M) sub-band features which are affected little by noise. These features should then be combined with the "and" operator. Without knowing where the noise occurs,

this subset may be *any* of the subsets of (N − M) sub-band features. This uncertainty can then be modelled with the "or" operator. Combining the two together we obtain a model for representing the useful information within the given feature set

$$o\vee = \bigvee_{n_1 n_2 \ldots n_{N-M}} o_{n1}\, o_{n2} \ldots o_{n_{N-M}} \tag{6}$$

where the "and" operator $\wedge$ between the o_n's has been omitted, and the "or" operator $\vee$ is taken over all possible combinations of N different values (1, ..., N) taken (N - M) at a time, giving a total of ${}^{N}C_{N-M}$ combinations. For example, in the simple case with four sub-bands, (6) can take one of the following four possible forms, corresponding to M = 0, 1, 2 and 3, respectively:

0) $o_1 o_2 o_3 o_4$

1) $o_1 o_2 o_3 \vee o_1 o_2 o_4 \vee o_1 o_3 o_4 \vee o_2 o_3 o_4$

2) $o_1 o_2 \vee o_1 o_3 \vee o_1 o_4 \vee o_2 o_3 \vee o_2 o_4 \vee o_3 o_4$

3) $o_1 \vee o_2 \vee o_3 \vee o_4$

Forms 0 and 3 correspond to the product model (1) and the general union model (3), respectively, and forms 1 and 2 correspond to the assumptions that there is one and two noisy sub-bands, respectively. In form 1, for example, the union of the four conjunctions will include one conjunction providing the joint probability of all three clean sub-bands; the other three conjunctions each contain a noisy

sub-band, with a correspondingly low probability on the correct model, and therefore make only a small contribution to the union probability associated with the correct model. In a similar way, in form 2 assuming two noisy sub-bands, one of the six conjunctions will correspond to the remaining two clean sub-bands and this conjunction will dominate the union probability associated with the correct model.

For convenience, we call (6) a *union model of order M*. As indicated above, the value of M corresponds to the maximum number of noisy sub-bands that can be accommodated in the model, in terms of leaving at least one conjunction consisting of only clean sub-bands. The product model (1), which includes a full conjunction of the sub-bands, corresponds to a union model with order $M = 0$ and therefore is best suitable for clean utterance without band corruption. The general union model (3) has an order $M = N - 1$, and thus may accommodate up to $N - 1$ noisy bands. Note that while a match between the order of the model and the number of noisy bands is desirable to maximise the information being extracted, a union model with order M may also be suited to situations where the number of noisy sub-bands is less than M.

For example, the above form 2, with order $M = 2$, may also be used to accommodate one noisy sub-band or none. This offers robustness against uncertainty on the number of corrupted bands. This characteristic

has been exploited previously for the selection of the model order, to seek a balance between the maximum performance and robustness. Details of this will be discussed later, along with a new algorithm for automatic order selection.

The expression for the union probability of (6) can be readily derived with o_n in (5) replaced by the appropriate conjunctions of sub-band features, i.e. $o_{n1} o_{n2} \ldots o_{nN-M}$, assuming independence between the features. This computation requires only the probability distributions of the individual sub-bands, as was required in the general union model discussed in the previous section.

IMPLEMENTATION

In this section, we first describe the implementation of the union model within a HMM framework, and then we describe the algorithms proposed for order selection.

A. *Incorporation into HMM*

We have built the above union model (6) into an HMM for combining the sub-band features at the *frame* level. Assume that there are N sub-bands, and that a speech utterance in each sub-band is represented by a sequence of frame vectors $o_n(1)$, $o_n(2)$, ... $o_n(T)$, $n = 1, \ldots, N$.

Combining the sub-band features at the frame level means that the union model (6) is applied at every frame time t, to combine the frame vectors $o_1(t)$, $o_2(t)$, ... $o_N(t)$ from all the sub-bands to obtain a union observation $O\vee(t)$, $t = 1, \ldots, T$. Then we modify the conventional HMM for this new observation sequence, by using a union-based observation probability distribution for each $O\vee(t)$. This HMM can be written as

$$P(O|\lambda) = \sum_S P(S|\lambda) \prod_{t=1}^{T} B_{s_t}(O_\vee(t)) \tag{7}$$

where O represents the frame sequence for all the sub-bands, $P(S|\lambda)$ is the probability of the state sequence S, and $B_i(O\vee)$ is the union based frame-level observation probability distribution in state i. As usual, the parameter set of the model, λ, includes the state transition probability matrix and initial state probability vector, which are needed for calculating the probability $P(S|\lambda)$ and the observation distribution set $\{Bi(O\vee)\}$. As described above with the assumption that the sub-band frames are mutually independent, the probability $B_i(O_\vee)$ is only a function of the individual probabilities $Bi(o_n)$'s where $Bi(o_n)$ represents the observation probability of the frame in sub-band n and state i. For a discrete-observation HMM, these sub-band observation probability distributions are readily available, and so $B_i(O\vee)$ can be readily calculated

by using the algorithm described above. However, note that (4) or (5), for computing the union probability, apply only to probabilities, not to probability densities or likelihoods. Therefore a special treatment is needed to resolve this issue when implementing the union model for a continuous-observation HMM, which employs an observation probability density $b_i(o_n)$ to account for the frame in sub-band n and state i. Basically, we seek an approximated probability based on a likelihood. However, this approximation is not needed in the model training stage, if the model is trained on clean speech data. Although $B_i(O\vee)$ varies with the order M for recognition, there is only one form, with order $M = 0$, that best matches a clean observation. Therefore in the training stage we can compute the union observation probability $B_i(O\vee)$ as the full conjunction probability $B_i(o_1)...B_i(o_N)$[1]. Since this probability is proportional to the likelihood $b_i(o_1)...b_i(o_N)$, we can train the model by maximising the likelihood function

$$p(O|\lambda) = \sum_S P(S|\lambda)\prod_{t=1}^{T}\prod_{n=1}^{N} b_{s_t}(o_n(t)) \qquad (8)$$

[1] More rigorously, the probability of a continuous o_n should be written as $B_i(x \in \Omega_n)$ i.e. the probability of a continuous random vector x falling into a sub-space Ω_n

surrounding o_n. But for simplicity we will keep using the expression $B_i(o_n)$.

and this can be accomplished by using the standard forward-backward re-estimation algorithm. In recognition, decisions are made by comparing the probability $P(O|\lambda)$, defined in (7), between different models. As with the conventional HMM, this probability can be computed by using the Viterbi algorithm, i.e.

$$\delta_t(j) = \max_i (\delta_{t-1}(i) + \log a_{ij}) + \log B_j(O \vee (t)) \tag{9}$$

where $\delta_t(i)$ is the log probability associated with a best state-sequence ending in state i for the observation up to time t, and a_{ij} is the state transition probability. With order $M \neq 0$, there may be two ways to obtain an approximated union probability $B_i(O\vee)$, based on the sub-band frame likelihoods $b_i(o_1),...,b_i(O_N)$. One way is to leave out the product term in (5), assuming that it is small and can be neglected in comparison to the other two additive terms. As such, the union probability $B_i(O_\vee)$ with $O\vee$ defined by (6) can be written as

$$\begin{aligned} B_i(O_\vee) &\cong \sum_{n_1 n_2 \ldots n_{N-M}} B(o_{n_1}) B_i(o_{n_2}) \ldots B_i(o_{n_{N-M}}) \\ &\propto \sum_{n_1 n_2 \ldots n_{N-M}} b_i(o_{n_1}) b_i(o_{n_2}) \ldots b_i(o_{n_{N-M}}) \end{aligned} \tag{10}$$

where the summation is over all possible combinations of N different values $(1, \ldots, N)$ taken $(N - M)$ at a time. Therefore (10) indicates a likelihood that may be used to approximate the union probability.

Alternatively, a sigmoid function may be used to approximate the sub-band frame probability $B_i(o_n)$ based on the likelihood $b_i(o_n)$, i.e.

$$B_i(o_n) \cong \frac{1}{1+e^{-\ln b_i(o_n)}} \tag{11}$$

This has the property that it produces an approximated probability that is proportional to the likelihood value, and at the same time satisfies the constraint $0 \leq B\,(o_n) < 1$ (this is required by (5) not to produce a negative probability). The probability $B_i(O\vee)$ with each $B_i(o_n)$ defined by (11) can thus be computed based on (5), including the product term. Because this term is usually very small (particularly for models with an order $M << N$), the two methods described above are based on (10) and (11) have been found to produce almost identical results.

Based on the assumption that the conjunction including only the clean bands should dominate the union probability for the correct model, (10) may be further approximated as

$$B_i(O_v) \cong \max_{n_1 n_2 \ldots n_{N-M}} b_i(o_{n_1}) b_i(o_{n_2}) \ldots b_i(o_{n_{N-M}}) \tag{12}$$

where the maximisation is over all possible combinations of N different values $(1,\ldots N)$ taken $(N - M)$ at a time. We have found in our experiments that, given the same order M ($M > 0$), the recognition results base on (10) and (12) are similar for low SNR conditions. However, in high SNR conditions, (10) was usually found to perform significantly better than (12). This is because (10) does not physically remove any sub-bands from recognition which (12) does. In high SNR conditions, those bands thrown away in (12) may still carry useful information.

B. Algorithms for order selection

A second embodiment of the present invention enables selection of an appropriate order to accommodate the corrupted sub-bands within an observation. As indicated above if there is no knowledge on the corrupting noise, it is safer to select a high order to accommodate as much noise as possible. However, because a higher-than-needed order will usually cause a loss of information due to unnecessary disjunction of the clean sub-bands, the order must be subject, for example, to an acceptable performance for clean speech recognition. We call this the balance fixed-order algorithm, which has been tested previously and has shown a limited success. In the following we describe an improved

algorithm, which derives the order automatically based on an optimality criterion.

As discussed above, an overestimated order (i.e. an order larger than the actual number of corrupted sub-bands) will lead to an unnecessary disjunction between the clean bands. This can cause some of the information relating to the joint probability distribution of the clean bands to be lost. On the other hand, an underestimated order (i.e. an order smaller than the actual number of corrupted sub-bands) will cause every conjunction in the union model to include, and so to be affected by, one or more corrupted sub-bands. Formally, we define the matched order as the order that equals the number of corrupted sub-bands. With this order, the union model will include a conjunction which contains all of the clean sub-bands together and no others, thereby capturing more discriminative information than either of the order-overestimated model or order-underestimated model, i.e. the order mismatched model. Because the order-matched model captures more clean band information, it should have more characteristics of a clean utterance than the order-mismatched model. This assumption forms the basis of our order selection algorithm. In particular, we use the state duration probability for clean utterance to estimate the matched order.

The state duration probability $P_i^u(d)$, for d frames in state i of phonetic unit u, is estimated in the training stage using the clean training data. Given

training stage using the clean training data. Given a test utterance, we perform recognition by using a set of union models, each with a different order, assuming that these will include the matched order. For each order, we obtain a recognition result (in the form of a unit sequence) $U(r) = u_1(r)u_2(r)...u_{n_r}(r)$ where r is the order index, along with the associated state duration $d_i(r)$, for each state i of $U(r)$. Because the model with the matched order captures the maximum clean band information, its state duration should be most similar to the state duration of a clean utterance. Therefore an appropriate estimate of the matched order would be the order whose associated state duration has the maximum probability, i.e.

$$\hat{r} = \arg\max_r \frac{1}{S(r)} \sum_{u \in U(r)} \sum_{i \in u} \ln P_i^u(d_i(r)) \qquad (13)$$

where $S(r)$ stands for the total number of states in $U(r)$. The final recognition result is then given by $U(\hat{r})$.

EXPERIMENTS

The TIDIGITS connected digit database was used to evaluate the performance of the new union model. This database contained connected digit strings from 225 adult speakers, conveniently divided into training and testing sets. The testing set contained a total of 6196 utterances from 113

speakers, each speaker contributing five utterances, containing 2, 3, 4, 5 and 7 digits, respectively. In recognition we assumed no advance knowledge of the number of digits in an utterance.

The speech was sampled at 8 kHz, and was divided into frames of 256 samples, with a between-frame overlap of 128 samples. For each frame, we used a mel-scaled filter bank to estimate the log-amplitude spectra of speech. Based on these log filter-bank spectra, both the full-band features and sub-bands features were calculated. The full-band features were used for comparison, which were the full-band MFCCs (mel-frequency cepstral coefficients) and were obtained by taking a DCT over the complete set of the log filter-bank spectra. The sub-band features were obtained by first grouping the filter-bank channels uniformly into sub-bands, and then, for each sub-band, performing a DCT for the log filter-bank spectra within that sub-band. This gives the sub-band MFCCs. In both cases, the first-order delta MFCCs were included in the feature vectors. The division of the speech frequency-band into sub-bands remains a topic of research. To effectively isolate any local frequency corruption from the other usable bands, a fine subdivision may be desirable. However, breaking the available frequency-band into too many independent sub-bands will cause much of the spectral dependency to be ignored, thus giving a poor phonetic discrimination. As an experimental study, we have tested the division of the available frequency-band into 3, 5

and 7 sub-bands, respectively, earlier for the E-set word recognition and now for the connected digit recognition. Both experiments indicate that the 5-band model appears to be a better choice in terms of the balance between the noise localisation and phonetic discrimination. Therefore in the following we focus on the experiments with five sub-bands (i.e. $N = 5$, in models (2) and (6)).

Specifically, these five sub-bands were grouped from a mel-scaled filter bank with 30 channels, each sub-band thus containing six log filter-bank spectral components for a frame. From these six components three MFCCs were derived, plus the delta parameters, as the feature vector of a sub-band frame. Thus, for this 5-band system, the overall size of the feature vector for a frame is 5 x 6 = 30. The full-band feature vector of a frame includes 20 components (10 MFCCs and 10 delta MFCCs), derived from a mel-scaled filter bank with 20 channels.

In addition to the union model, for comparison, we also implemented a baseline HMM which used the above full-band features and a product model which is a special case of the union model with order $M = 0$. All these models were based on Gaussian mixture densities with diagonal covariance matrices, and were trained on clean training data. In particular, each digit was modelled with 10 states, and a silence model with one state was built to account for the silences surrounding each utterance and the optional silences between digits. Each of these

states contained eight mixtures. For the union model, we also recorded the histograms of state occupancy occurring in each digit, as the estimates of the state duration probabilities. The state duration probability was used only for selecting the model order, as described above and was not incorporated into the HMMs for scoring the observations.

In the following we first present the recognition results by the union model under various testing conditions. Then we discuss its generalisation to the combination of different types of feature streams, and its combination with a conventional noise-reduction technique.

A. Tests with clean speech

Table I presents the string accuracy obtained by the union model and the baseline model, respectively, for clean utterance recognition. As shown in the table, our baseline HMM achieved a string accuracy of 97.53%,

Based on (6), for a union model with N sub-bands (now $N = 5$), recognition can be performed with different orders (i.e. M) within the range $0 \le M \le N-1$ (now $0 \le M \le 4$). Table I presents the accuracy obtained by using each of these individual orders, along with the accuracy based on the automatically selected order. Note that at order 0, the union model is equivalent to a product model.

As described earlier, since there is no band corruption, a clean speech utterance is better characterised by a full conjunction of all the sub-band features. This explains why the product model, derived from such a conjunction, produced the best performance among all the orders within the range $0 \le M \le 4$. As expected, the performance of the union model decreased as the order was increased, because of the disjunction between the clean sub-band features.

Given a test utterance, the above models with fixed orders each produced a recognition result, tagged by the associated order. The automatic order selection algorithm, (12), was then applied to these results to select an order with maximum state duration probability, thereby obtaining the final recognition result. As shown in Table I, this gives an accuracy that is very close to the accuracy obtained by the best (i.e. matched) order - order 0. Fig. 1 shows the histograms of the orders selected by the algorithm. As indicated in Fig. 1(a), for clean test utterances, the algorithm correctly selected more than 50% of the orders. This correct selection rate may be improved by putting a restriction on the order range searched by the algorithm. For example, we tested the use of a smaller range $0 \le M \le 3$ instead of $0 \le M \le 4$ and ended with slightly better result for clean utterance recognition. However, allowing the uncertainty of the environment, in the following all automatic orders were selected from the order range $0 \le M \le 4$.

B. Tests with stationary band-selective noise

To evaluate the robustness of the union model, we first tested the model for the utterances corrupted by stationary band-selective noise. The noise, added to the speech, was generated by passing Gaussian white noise through a band-pass filter with a 3-dB cut-off bandwidth of 100 Hz and a varying central frequency. In particular, six different central frequencies were considered, these were 600 Hz, 850 Hz, 1200 Hz, 1500 Hz, 2000 Hz and 2500 Hz. These were chosen to create the effects that there were one sub-band, two sub-band and three sub-band corruptions, respectively, within the five sub-bands of the system. Specifically, the noises with central frequencies 600 Hz, 1200 Hz and 2000 Hz were located within sub-band 2, 3 and 4, respectively, and each thus caused only one sub-band corruption; the noises with central frequencies 850 Hz, 1500 Hz and 2500 Hz were located around the border of sub-bands 2 and 3, 3 and 4, and 4 and 5, respectively, and each thus caused two sub-band corruptions. The noises corrupting three sub-bands were generated by combining two noise components with different central frequencies, in particular, 600 Hz and 1500 Hz (corrupting sub-bands 2, 3 and 4), and 1200 Hz and 2500 Hz (corrupting sub-bands 3, 4 and 5), respectively. The six band-selective noises, plus the two combined noises, resulted in a total of eight different noise conditions. For all conditions, we assumed no prior knowledge of the noise being available for the union model.

Table II presents the recognition results, as a function of the number of corrupted sub-bands and SNR within each test utterance. These results are averaged over the appropriate noise conditions producing the same number of noisy sub-bands, as elaborated above. From Table II, two particularly useful observations can be made for the union model. Firstly, for each given SNR condition, the fixed-order model achieved the maximum accuracy at the order that matched the number of corrupted sub-bands. Secondly, the automatic-order model was able to achieve an accuracy that was close to the matched-order accuracy, throughout all test conditions. In particular, we see that in two cases (with three noisy bands, SNR=10 dB and 5 dB, respectively) the automatic-order model achieved a higher recognition accuracy than the corresponding matched-order accuracy (i.e., 76.27% vs 72.32%, and 64.81% vs 64.75%, respectively). This may be because the order selection algorithm is operated on each utterance basis, so it may choose an order which includes some noisy bands, in which the local SNRs are high. Fig. 1(b)-(d) show the histograms of the orders selected by the algorithm for the noisy conditions. We see that in each condition, the algorithm selected the matched order with the highest frequency. Based on Tables I and II, we then may conclude that, equipped with the automatic order selection algorithm, the union model can effectively achieve a near matched-order performance for both clean and noisy conditions, without requiring any information on the nature of the

environment (i.e. clean or noisy) and on the noise (i.e. the location and number of noisy sub-bands), if the environment is noisy.

We next conducted comparisons between the union model with automatic order and hence requiring no knowledge on the noise, with two other models with knowledge on the noise. The first model we compared was an ideal missing-feature model, or the "oracle" model which assumed a full a priori knowledge of the corrupted sub-bands and removed those bands manually from the recognition. The second model being compared was a baseline HMM equipped with a Wiener filtering front-end for removing the noise, based on the assumption that the noise was stationary and for which a spectral estimate was available. The spectrum of the stationary band-selective noise was estimated in the interval without speech. The spectral estimate was then used to build a Wiener filter, derived from spectral subtraction to enhance the noisy signal before recognition. Table III presents the results. As expected, the oracle model performed better than the union model, and the gap between their performances is significant in many cases. Later we will discuss an improvement over the union model, to reduce this performance gap. In one case, with three noisy sub-bands and SNR=10 dB, the union model outperformed the oracle model. This is because throwing away the three bands with relatively high SNR in the oracle model caused a loss of much useful information. However, when all these bands were included, it gave an accuracy of

only 28.18%, as shown in Table I. So a "soft" rather than a binary decision is preferred as to whether to include or exclude a particular sub-band. The union model provides such a soft-decision mechanism. It is capable of ignoring those noisy bands that significantly violate the statistics of the training data population; but it physically removes no band from recognition, as such each band retains a contribution, proportional to its likelihood value, to recognition. Comparing Table III with Table II, we see that the Wiener filtering considerably improved the performance of the baseline model. However, the union model still performed significantly better than the baseline model with Wiener filtering, throughout all test conditions.

C. Test with real-world, non-stationary noise

Next, we tested the union model, with automatic order, for recognising utterances corrupted by some real-world noises. The noise data used in the experiments are shown in Fig. 2, which include the sounds of a ding, a telephone ring, a whistle, which were extracted from the sound files "ding.wav", "ring.wav" and "whistle.wav", respectively, from the Windows operating system, and the sounds of "contact" and "connect", which were used in an internet tool (ICQ) for on-line contact, chat and sending messages. These noises each have a dominant band-selective characteristic, and the noises "contact" and "connect" are particularly non-

stationery. These noises were added, respectively, to each of the test utterances for recognition experiments. Table IV presents the string accuracy obtained for each of these noises and the average accuracy over all these noises. As a reference, Table IV also includes the results given by the baseline model. No noise reduction technique was employed in the baseline model, due to the non-stationary nature of the noise and due to the assumption that there was no prior knowledge about the noise.

Table IV indicates that the performance of the union model for the telephone-ring noise and "connect" noise is less significant in comparison to the performance for the other three types of noise. This is because both the telephone-ring noise and "connect" noise have particular multi-band characteristics. For the telephone-ring noise, for example, the first two tones lay in bands 3 and 4, respectively, and the last two tones fell within band 5, which thus affected 3 sub-bands. We have experienced weakness of the sub-band method for dealing with wide-band noise. Wide-band noise affects all sub-bands, which therefore violates the noise-localization assumption made in the sub-band model. For a system to be capable of dealing with both narrow-band and wide-band noises, a combination of different techniques may be needed. We will show such an example later.

D. Generalisation to partial feature stream corruption

So far we have described a union model for extracting useful features from a set of sub-band feature streams $\{o_1, o_2 \ldots, o_N\}$, where each o_n represents the feature stream of a specific sub-band. In a third embodiment of the present invention, this model may be generalised by considering the feature set $\{o_1, o_2, \ldots, o_N\}$, as a collection of more types of feature stream rather than only the sub-band feature stream. In speech recognition, a speech utterance may be represented by multiple feature streams, typically, the static spectra and dynamic spectra, over varying time scales. In real-world applications, due to the background noise or channel effects, there may be only a subset of the given feature streams that remain reliable. For example, the static spectral features are more sensitive to a stationary or slowly-varying noise than the dynamic spectral features. If a feature stream is adversely affected, it should play a less significant role than the other unaffected streams in recognition. However, without prior knowledge of the environmental or noise condition, it can be difficult to decide which subset of the feature streams provides reliable information. This uncertainty may be dealt with by using the union model. For this, we rephrase the above sub-band combination problem as a general feature selection problem, i.e. selecting reliable features from a

feature set $\{o_1, o_2 \ldots, o_N\}$, where each o_n represents a specific feature stream, given that some of the o_n's may be corrupted, but without knowledge about their identity.

As an application, we have generalised our previous sub-band union model by applying the union not only to the combination of the sub-bands, but also to the combination of the static and dynamic feature streams, to further select the feature stream within each sub-band that is least affected by noise. Specifically, we separated the static feature and dynamic feature within each sub-band into two feature streams o_n and Δo_n, where Δo_n represents the dynamic feature stream (i.e. ΔMFCCs), and then we modelled the entire feature set $\{o_{1,} \ldots, o_N, \Delta o_1, \ldots, \Delta o_N\}$ with a union model with $2N$ input streams and a full order range $0 \le M \le 2N-1$. With the previously defined 5-band system, we then had a union model with 10 input feature streams (five for MFCCs and five for ΔMFCCs, each consisting of 3 components for each frame) and a full range order $0 \le M \le 9$. Using this generalised union model, we repeated all the previous experiments under exactly the same test conditions. The generalised model used automatic orders selected from an order range $0 \le M \le 8$.

Tables V and VI present the string accuracy obtained by the generalised union model, along with the average error reduction in comparison to the previous union model without applying the union for

the static and dynamic feature streams, as shown in Tables I, III and IV. Comparing Table V with Table I, we see that the generalised union model even improved the accuracy for clean utterance recognition. Comparing Table V with Table III, for stationary band-selective noise, we see that the generalised model significantly improved over the previous union model for all noise conditions, particularly for the conditions with multiple noisy bands. Comparing Table V with the oracle model in Table III, we see that the generalised union model outperformed the oracle model in many cases, and it actually achieved better average performance than the oracle model. Table VI shows the string accuracy by the generalised union model in real-world, non-stationary noise, corresponding to Table IV. Comparing these two tables, we again see that the generalised union model significantly improved the accuracy for all noise conditions. Improvements for the noisy cases may be due to the separation and removal of those static features that were more adversely affected by the noise.

E. *Combination of Techniques*

So far we have assumed no prior knowledge about the noise. This is typical for some random, abrupt noises. However, real-world noise may be a mixture of stationary noise and abrupt noise. For stationary noise, with reasonably sufficient observations, it is possible to obtain an estimate of the noise characteristics. In a fourth embodiment of the present invention, we consider the

building of a system in which the union model and some conventional noise-reduction techniques are combined, to deal with this type of mixed noise. The stationary noise component may be removed, for example, by spectral subtraction for additive noise, or by cepstral mean subtraction for convolutive noise. The remaining unknown unexpected noise component can be dealt with by the union model if it has a band-selective characteristic.

We have tested such a system by creating noisy speech data involving both stationery noise and unknown, band-selective noise, both being additive. Specifically, the stationary noise was a car noise, obtained from the Aurora 2 database, which exhibited a wide-band characteristic; the band-selective noise was a whistle, as shown in Fig. 2, which simulated a further unknown and unexpected band-selective corruption occurring to the utterance. To reduce the stationary noise component, we may use the Wiener filtering technique as described above. Here we considered a different technique, i.e. noise compensation. In particular, we assumed that we had the models trained in the car environment, so that the mismatch between the model and data, due to the existence of the stationary noise, could be reduced. While we assumed knowledge about the occurrence of the stationary noise, we assumed no knowledge about the occurrence of the whistle during the utterance. The SNR's of the two noise components were calculated separately relative to the clean speech data, and each was 10 dB (so the overall SNR within

each utterance was about 7 dB). The generalised union model described above was used in this experiment. Table VII presents the recognition results, showing the advantage of the combination of the union model and noise compensation technique for dealing with the mixed noise.

We then further developed this combination into a simple parallel-environment model, in which two sets of generalised union models, trained for clean condition and car condition respectively, were run in parallel, and the final result was selected using the order selection algorithm over the two sets of models. This model removes the requirement for a knowledge of the environment (i.e. clean or car). For clean speech input, this model produced a string accuracy of 95.30%, and for the noisy speech input, assuming the same mixed noise as described above, this model produced a string accuracy of 74.66%. Both accuracies were close to their respective environment-model matched accuracy, i.e. 96.21% and 75.21%, shown in Table V and Table VII, respectively.

It will be appreciated that various improvements and modifications can be made without departing from the scope of the invention.

Whilst the invention has been described with specific embodiments relating to speech recognition, it will be appreciated that the invention is applicable to any other areas of signal processing

and pattern recognition involving partial unknown feature corruption, for example, image processing, statistical language processing, communication, and artificial intelligence.

Alternative techniques for dealing with known or trainable noise or environmental effects may be incorporated into the invention, for example, speaker adaptation to accommodate speaker variation, or recognition of key words.

In the context of speech recognition, the principle of the invention can be extended to the combination of units at a higher level, for example phoneme or syllable.

TABLE I

STRING ACCURACY (%) FOR CLEAN UTTERANCES, FOR THE UNION MODEL WITH FIXED ORDERS AND AUTOMATICALLY SELECTED ORDER (AO), AND FOR THE BASELINE HMM. AT ORDER 0, THE UNION MODEL IS EQUIVALENT TO A PRODUCT MODEL

Union Model						Baseline HMM
Order						
0	1	2	3	4	AO	
96.48	95.08	92.03	86.99	64.11	95.58	97.53

TABLE II

STRING ACCURACY (%) IN STATIONARY BAND-SELECTIVE NOISE, FOR THE UNION MODEL WITH FIXED ORDERS AND AUTOMATICALLY SELECTED ORDER (AO), AND FOR THE BASELINE HMM. THE MATCHED-ORDER ACCURACY FOR THE UNION MODEL IS SHOWN IN ITALIC

SNR (dB)	# Corrupted Bands	Union Model Order 0	1	2	3	4	AO	Baseline HMM
10	1	58.04	*92.81*	89.92	81.52	52.93	90.67	61.62
	2	47.33	76.47	*88.65*	79.11	46.85	86.63	63.16
	3	28.18	59.74	72.13	*72.32*	42.88	76.27	34.20
5	1	40.98	*90.60*	87.24	76.77	46.64	88.29	37.04
	2	31.04	61.10	86.82	76.10	42.87	83.91	38.85
	3	9.35	35.55	53.50	*64.75*	37.10	64.81	13.66
0	1	24.35	*85.33*	82.33	69.38	37.58	83.93	17.70
	2	20.05	42.94	83.95	71.35	38.20	79.89	20.32
	3	2.86	20.27	34.47	*56.57*	31.58	53.95	3.77

TABLE III

COMPARISONS OF STRING ACCURACY (%) IN STATIONARY BAND-SELECTIVE NOISE, FOR THE UNION MODEL WITH AUTOMATIC ORDER, FOR THE ORACLE MODEL WITH A FULL A PRIORI KNOWLEDGE OF THE NOISY BANDS, AND FOR THE BASELINE HMM WITH WIENER FILTERING (WF)

SNR (dB)	Model	#Corrupted Bands			Average
		1	2	3	
10	Union	90.67	86.63	76.27	84.52
	Oracle	94.31	89.65	66.73	83.56
	Baseline (WF)	79.61	81.81	64.42	75.28
5	Union	88.29	83.91	64.81	79.00
	Oracle	93.21	88.39	65.18	82.26
	Baseline (WF)	60.47	62.15	36.95	53.19
0	Union	83.93	79.89	53.95	72.59
	Oracle	89.83	86.46	62.85	79.71
	Baseline (WF)	29.13	36.76	14.61	26.83

TABLE IV

STRING ACCURACY (%) IN REAL-WORLD NON-STATIONARY NOISE, FOR THE UNION MODEL WITH AUTOMATIC ORDER, AND FOR THE BASELINE HMM

SNR (dB)	Model	Noise Type					Average
		Ding	Tel Ring	Whistle	Contact	Connect	
10	Union	85.30	72.85	88.62	87.41	74.13	81.66
	Baseline	65.28	60.23	50.44	53.62	41.59	54.23
5	Union	80.46	60.77	86.06	84.60	58.76	74.13
	Baseline	43.56	34.49	25.87	30.57	16.03	30.10
0	Union	75.02	50.81	82.18	79.62	36.73	64.87
	Baseline	22.26	17.75	8.28	14.27	4.50	13.41

TABLE V

STRING ACCURACY (%) FOR CLEAN SPEECH AND IN STATIONARY BAND-SELECTIVE NOISE, FOR THE GENERALISED UNION MODEL, AND AVERAGE ERROR REDUCTION (%) IN COMPARISON TO THE PREVIOUS UNION MODEL IN TABLES I AND III, ALL WITH AUTOMATIC ORDERS

SNR (dB)	# Corrupted Bands			Average	Ave. Error Reduction
	1	2	3		
Clean	96.21				14.25
10	92.49	90.75	86.45	89.90	34.75
5	90.55	88.06	80.38	86.33	34.90
0	87.10	84.65	70.97	80.91	30.35

TABLE VI

STRING ACCURACY (%) IN REAL-WORLD NON-STATIONARY NOISE, FOR THE GENERALISED UNION MODEL, AND AVERAGE ERROR REDUCTION (%) IN COMPARISON TO THE PREVIOUS UNION MODEL IN TABLE IV, BOTH WITH AUTOMATIC ORDERS

SNR (dB)	Noise Type					Average	Ave. Error Reduction
	Ding	Tel Ring	Whistle	Contact	Connect		
10	90.96	81.99	90.95	88.15	79.21	86.25	25.02
5	88.12	73.87	88.75	85.31	65.80	80.37	24.12
0	84.68	62.90	84.88	81.81	44.96	71.85	19.86

TABLE VII

STRING ACCURACY (%) IN MIXED STATIONARY WIDE-BAND NOISE (CAR) AND UNKNOWN BAND-SELECTIVE NOISE (WHISTLE), EACH WITH AN SNR=10 DB, SHOWING THE EFFECTIVENESS OF COMBINING THE NOISE COMPENSATION TECHNIQUE AND THE UNION MODEL

	No Noise Compensation	With Noise Compensation
Union	35.75	75.21
Baseline	35.93	56.55

CLAIMS

1. A method of interpreting features for signal processing and pattern recognition in which recognition of a signal or pattern is enabled by a model in which the sample to be interpreted is considered as a set of N observations, M of which are corrupt, and a disjunction is performed over all possible combinations of N different values (1,…,N) taken N-M at a time.

2. A method as claimed in Claim 1 wherein $0 < M \leq N-1$.

3. A method as claimed in either preceding Claim in which the value of M, namely the number of corrupt observations, defines an order of the model, and is estimated using an optimality criterion in which:

   it is assumed that the matched order is the order having the most characteristics of a clean signal,
   an aspect of the clean signal is selected,
   the values of the aspect are compared for different orders, and
   the chosen order is defined as the order for which the value of the aspect is closest to that of a clean signal.

4. A method as claimed in any preceding Claim wherein the signal to be processed is a speech signal.

5. A method as claimed in any preceding Claim wherein the set of N observations comprises a set of N sub-band feature streams.

6. A method as claimed in Claim 3 in which said selected aspect is a state duration probability.

7. A method as claimed in Claim 6 in which the optimality criterion is obtained from the order selection algorithm

$$\hat{r} = \arg\max_{r} \frac{1}{S(r)} \sum_{u \in U(r)} \sum_{i \in u} \ln P_i^u(d_i(r))$$

where: r is the order index;
$\hat{r}$ is the order index with the highest associated state duration probability;
U(r) is a recognition result;
S(r) stands for the total number of states in U(r);
$P_i^u(d_i(r))$ is the state duration probability for *d* frames in state *i* of phonetic unit *u*.

8. A method as claimed in any preceding Claim in combination with conventional signal filtering techniques which remove known stationary corruptions.

9. A method as claimed in any of the preceding Claims substantially as hereinbefore described with reference to the accompanying tables and drawings.

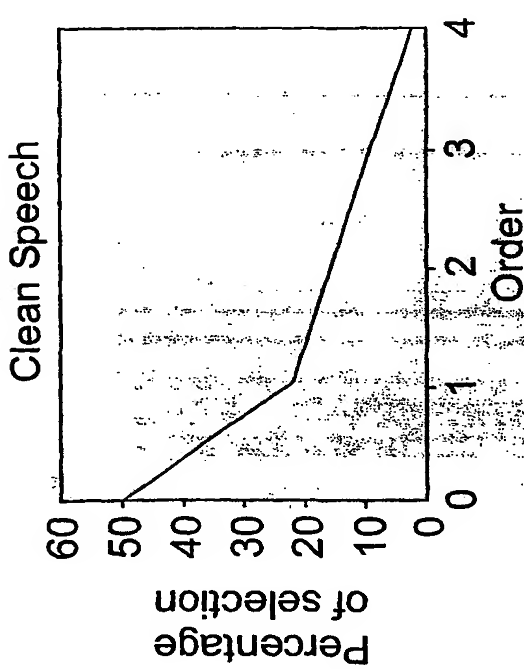

Fig. 1a

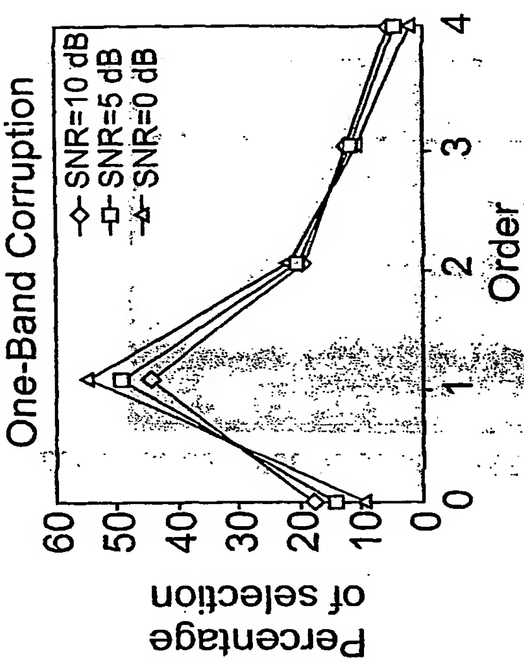

Fig. 1b

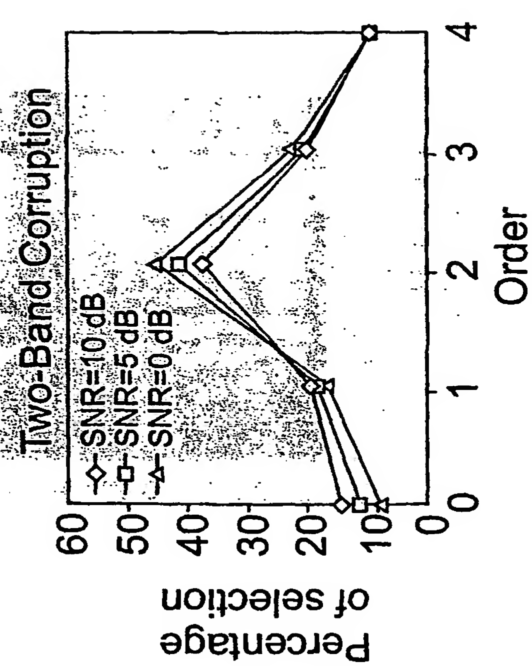

Fig. 1c

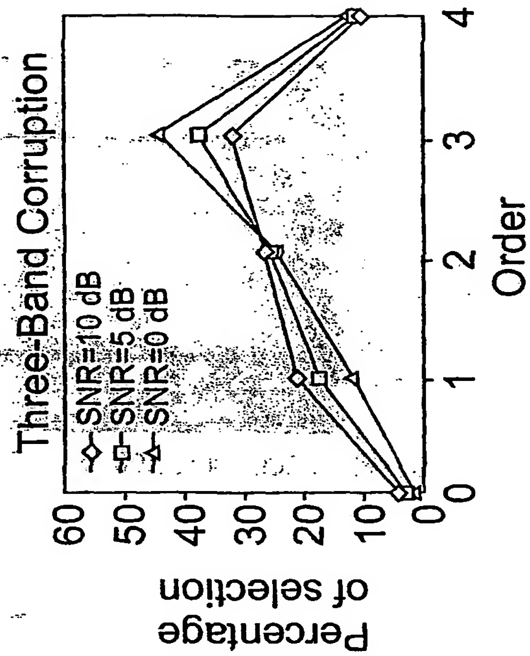

Fig. 1d

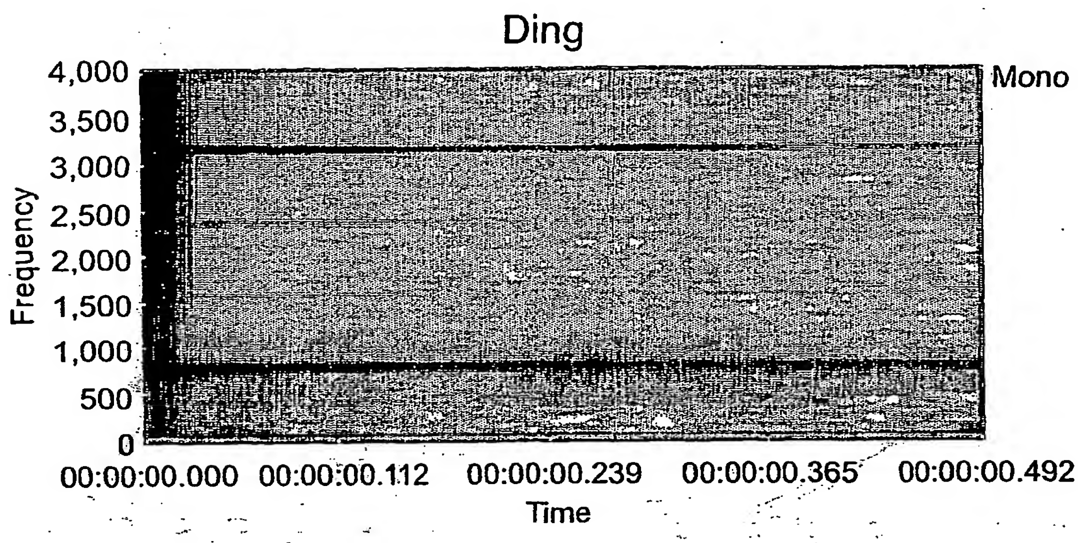
Ding
4,000
3,500
3,000
2,500
2,000
1,500
1,000
500
0
Frequency
Mono
00:00:00.000
00:00:00.112
00:00:00.239
00:00:00.365
00:00:00.492
Time

Fig. 2a

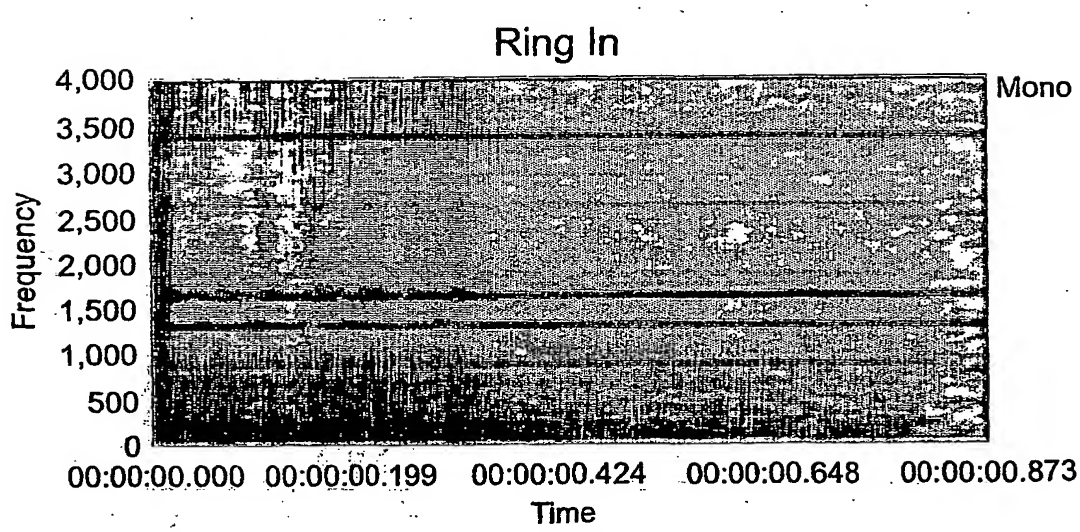
Ring In
4,000
3,500
3,000
2,500
2,000
1,500
1,000
500
0
Frequency
Mono
00:00:00.000
00:00:00.199
00:00:00.424
00:00:00.648
00:00:00.873
Time

Fig. 2b

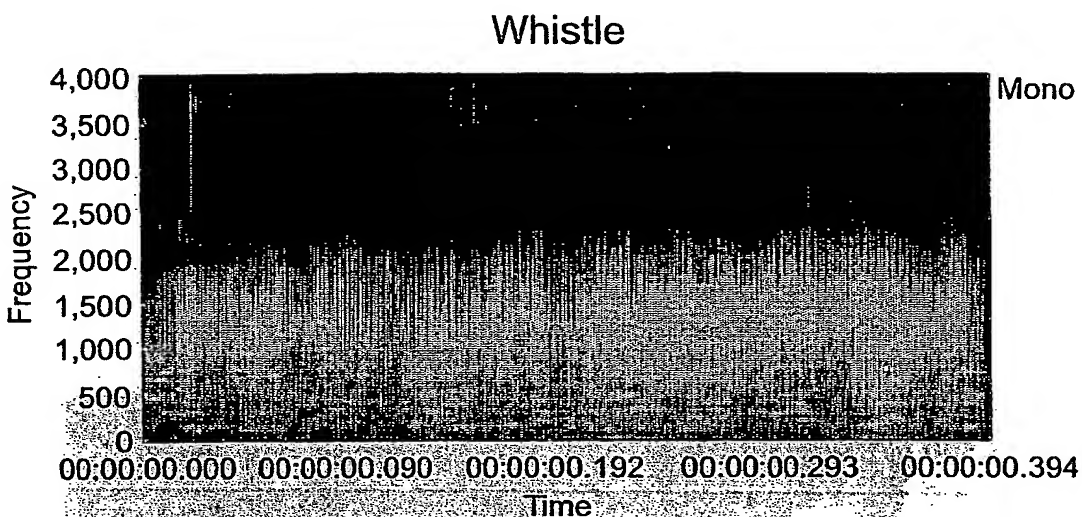
Whistle
4,000
3,500
3,000
2,500
2,000
1,500
1,000
500
0
Frequency
Mono
00:00:00.000
00:00:00.090
00:00:00.192
00:00:00.293
00:00:00.394
Time

Fig. 2c

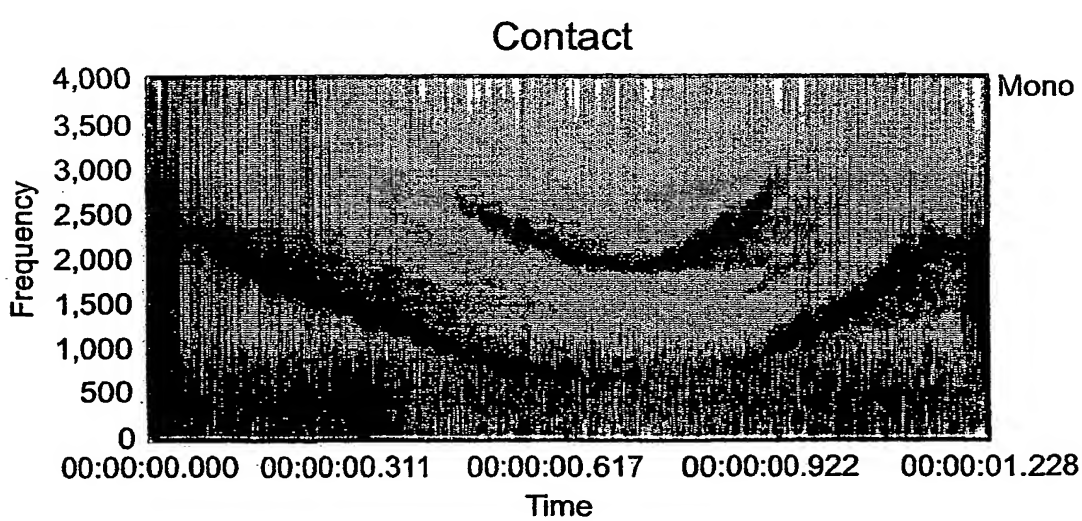
Contact
4,000
3,500
3,000
2,500
2,000
1,500
1,000
500
0
Frequency
Mono
00:00:00.000
00:00:00.311
00:00:00.617
00:00:00.922
00:00:01.228
Time

Fig. 2d

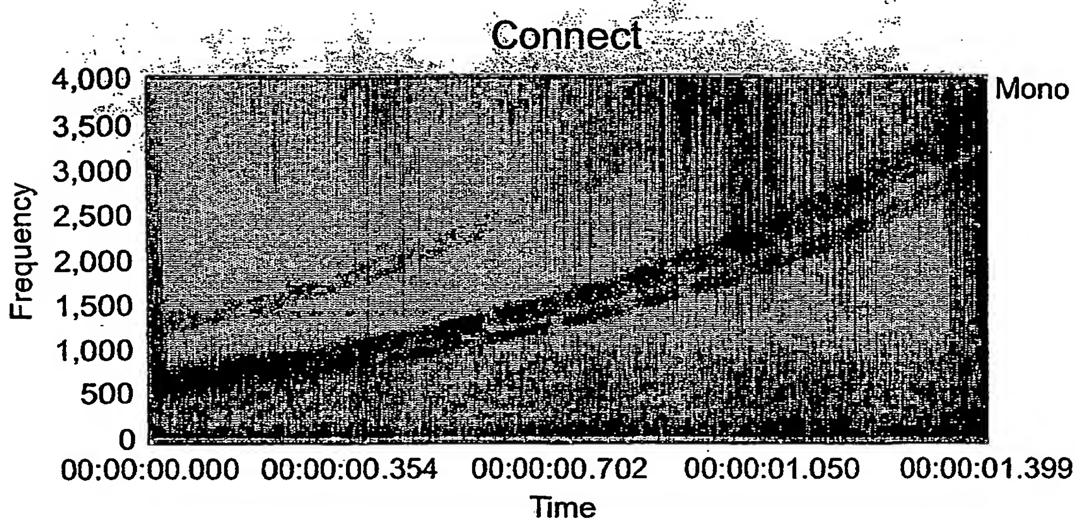
Connect
4,000
3,500
3,000
2,500
2,000
1,500
1,000
500
0
Frequency
Mono
00:00:00.000
00:00:00.354
00:00:00.702
00:00:01.050
00:00:01.399
Time

Fig. 2e

INTERNATIONAL SEARCH REPORT

PCT/GB 02/02197

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 G10L15/20

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 G10L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

EPO-Internal, WPI Data, INSPEC

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
P,X	JANCOVIC P ET AL: "A probabilistic union model with automatic order selection for noisy speech recognition" JOURNAL OF THE ACOUSTICAL SOCIETY OF AMERICA, SEPT. 2001, ACOUST. SOC. AMERICA THROUGH AIP, USA, vol. 110, no. 3, pages 1641-1648, XP001100608 ISSN: 0001-4966 the whole document --- -/--	1-8

[X] Further documents are listed in the continuation of box C.

[ ] Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
9 August 2002	22/08/2002
Name and mailing address of the ISA European Patent Office, P.B. 5818 Patentlaan 2 NL – 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Authorized officer Quélavoine, R

INTERNATIONAL SEARCH REPORT

PCT/GB 02/02197

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	JI MING ET AL: "Union: a new approach for combining sub-band observations for noisy speech recognition" SPEECH COMMUNICATION, APRIL 2001, ELSEVIER, NETHERLANDS, vol. 34, no. 1-2, pages 41-55, XP002209287 ISSN: 0167-6393 the whole document	1-5,8,9
X	JI MING ET AL: "A probabilistic union model for sub-band based robust speech recognition" 2000 IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING. PROCEEDINGS (CAT. NO.00CH37100), 5 - 9 June 2000, pages 1787-1790 vol.3, XP002209288 ISTANBUL, TURKEY, Piscataway, NJ, USA, IEEE, USA ISBN: 0-7803-6293-4 the whole document	1-5,8,9
A	JANCOVIC P ET AL: "Combining multi-band and frequency-filtering techniques for speech recognition in noisy environments" TEXT, SPEECH AND DIALOGUE. THIRD INTERNATIONAL WORKSHOP, TSD 2000. PROCEEDINGS (LECTURE NOTES IN ARTIFICIAL INTELLIGENCE VOL.1902), 13 - 16 September 2000, pages 265-270, XP008006658 BRNO, CZECH REPUBLIC, Berlin, Germany, Springer-Verlag, Germany ISBN: 3-540-41042-2 the whole document	1-9